* Regression : Y numerical $\quad Y_i = \alpha + \beta X_i + \epsilon_i$

* Logistic : $Y \sim \text{Ber}(p) \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i$

  ↳ or $Y \sim$ Multinomial

* Time Series : Y numerical but indexed by time (not iid)

  $Y_t = \alpha + \beta Y_{t-1}$

* Poission Reg : Y is count data

* Survival Reg : Cox Regression

* Bradley-Terry : Weighted Logistic Regression

* Bass Model : Forecasting newer launched products

* Lee Carter Model : Mortality forecasting

* Example : Bradley - Terry Model Pairwise comparison

There are 30 basket ball teams in the NBA, each playing 82 games in the regular season (Total games $\frac{30 \times 80}{2} = 1230$). We observe at the end of each season which two teams (i, j) played in each game and who won. How can we rank the teams and/or determine the strength of each team?

The simplest strategy is to compare the # of games won by each team however the NBA is structured so that every team plays every other team a different number of games (between

2 and 4) so the teams have different strength of schedule meaning that some teams play stronger opponents more frequently than do other teams. These teams might have worst win - loss records, in fact, be better than other teams them won more games against weaker opponents.

A model - based approach to solve this problem is the following:

Let $\beta_i \in \mathbb{R}$ represent the strength of team i. The outcome of the game played between (i, j) be determined by $\beta_i - \beta_j$.

BT Model treats this outcome as an independent bernoulli random variable with distribution $Ber(p_{ij})$, where log odds corresponding to probability p_{ij} that team i beats team j is modeled as:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_i - \beta_j \quad \Rightarrow \quad p_{ij} = \frac{e^{\beta_i - \beta_j}}{1 + e^{\beta_i - \beta_j}}$$

What if it was a home game?

If we always order each pair (i, j) so that team i is the home team and j is the away team, then we may incorporate a home - court advantage by including an intercept term α:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha + \beta_i - \beta_j \quad \Rightarrow \quad p_{ij} = \frac{e^{\alpha + \beta_i - \beta_j}}{1 + e^{\alpha + \beta_i - \beta_j}}$$

This increases the log odds of the home team winning in every game by a constant value α.

1) How do we estimate α, β_i, β_j?

α) Testing home advantage $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$

1) Suppose we observe n total games played between $(i_1, j_1), \dots, (i_n, j_n)$ between these K teams where each (i, j) is a pair of distinct teams in $\{1, 2, \dots, K\}$, we denote the home team is team i.

Let $Y_1, \dots, Y_n \in \{0, 1\}$ be such that $Y_m = \begin{cases} 1, & i_m \text{ beats } j_m \\ 0, & i_m \text{ loses } m^{th} \text{ game} \end{cases}$

The likelihood for the parameters $\theta = (\alpha, \beta_2, \dots, \beta_K)$

$$L(\alpha, \beta_2, \dots, \beta_K) = \prod_{m=1}^{n} P_{i_m, j_m}^{Y_m} \left(1 - P_{i_m, j_m}\right)^{1 - Y_m}$$

$$= \prod_{m=1}^{n} \left(\frac{P_{i_m, j_m}^{Y_m}}{(1 - P_{i_m, j_m})^{Y_m - 1}} \right)$$

$$\ell(\theta) = \sum_{m=1}^{n} Y_m \log(P_{i_m, j_m}) + \sum_{m=1}^{n} (1 - Y_m) \log(1 - P_{i_m, j_m})$$

$$= \sum_{m=1}^{n} Y_m \log\left(\frac{P_{i_m, j_m}}{1 - P_{i_m j_m}} \right) + \log(1 - P_{i_m, j_m})$$

$$= \sum_{m=1}^{n} Y_m (\alpha + \beta_i - \beta_j) - \log(1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}})$$

$$\frac{\partial \ell(\theta)}{\partial \alpha} = \sum_{m=1}^{n} Y_m - \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_i} = \sum_{m : i_m = 1} Y_m - \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}} + \sum_{m : j_m = 1} \left(-Y_m + \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}} \right)$$

The model $P_{ij} = \dfrac{e^{\beta_i - \beta_j}}{1 + e^{\beta_i - \beta_j}}$ is an over parametrized model

in the sense that, it is exactly the same if we add a fixed constant c to all β_i, the difference $\beta_i - \beta_j$ remains unchanged.

We may fix this problem by setting $\beta_i = 0$ for a particular team $\beta_i = 0$ then for every other team j, $\beta_j = \beta_j - 0$ represents the log odds that the team j beats the team i.

* Linear Models : $Y = X\beta + \epsilon \qquad Y \sim N(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2)$

$$L(\beta | x) = \prod_{i=1}^{n} f(X | \beta)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(\frac{-(Y-X\beta)^T(Y-X\beta)}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta}{2\sigma^2}\right)$$

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta)$$

$$\frac{\partial \ell}{\partial \beta}(\beta, \sigma^2) = \frac{1}{2\sigma^2}(-2X^TY + 2X^TX\beta) = 0 = X^TY - X^TX\beta$$

$$\frac{\partial \ell}{\partial \beta}(\beta, \sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}((Y-X\beta)^T(Y-X\beta)) = 0$$

$$\hat{\beta}_{MLE} = (X^TX)^{-1}X^TY$$

$$\hat{\sigma}_{MLE} = \frac{1}{n}(Y - X\hat{\beta}_{MLE})^T(Y - X\hat{\beta}_{MLE})$$

$$\frac{\partial \ell}{\partial \beta_m} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_{im}\left(Y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)$$

$$\frac{\partial^2}{\partial \beta_m \partial \beta_\ell} = \nabla^2 \ell(\beta) = \frac{-1}{\sigma^2}\sum_{i=1}^{n} x_{i\ell} \cdot x_{im} = \frac{-X^TX}{\sigma^2}$$

Hence, $I_Y(\beta) = \frac{-X^T X}{\sigma^2}$ so $\hat{\beta} \sim N(\beta, I_Y(\beta)^{-1}) = N(\beta, \sigma^2 (X^T X)^{-1})$

* Logistic Regression :

An internet company would like to understand what factors influence whether a visitor to a webpage on an advertisement. Suppose it has some historical data on n ad impressions, each impression corresponds to a single ad being shown to a single visitor. For the i^{th} impression, let $Y_i \in \{0,1\}$; $Y_i = 1$ if the visitor clicked on the ad, and $Y_i = 0$ otherwise. The company also has various attributes available for each impression, position of the ad, size of the ad, nature of the ad, age, gender, timing, month, etc $(x_{i1}, \dots, x_{ip})$. Predict target $Y_i \in \{0,1\}$. Binary classification problem.

The logistic regression model assumes each response Y_i as an independent random variable with bernoulli (p_i) distribution where the log odds corresponding to p_i is modelled as a linear combination of the covariates (and intercept term).

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$ without error terms as log odds is not a RV

↳ range $\mathbb{R}$ and $p_i \in [0,1]$

$$\frac{p_i}{1-p_i} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}$$

$$p_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}} = \mathbb{P}(Y_i = 1)$$

Interpretation of β_0 : Baseline log odds of the visitor clicking on the ad when all the covariates are zero.

Interpretation of β_1 : The amount of increase in log odds if the value of the covariate x_{ij} is increased by 1 unit.

$$L(\beta_0, \dots, \beta_p) = \prod_{i=1}^{n} f(Y_i \mid \beta_0, \dots, \beta_p)$$

$$= \prod_{i=1}^{n} p_i^{Y_i} \cdot (1 - p_i)^{1 - Y_i}$$

$$\ell(\beta_0, \dots, \beta_p) = \sum_{i=1}^{n} Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)$$

$$= \sum_{i=1}^{n} Y_i \left(\log(p_i) - \log(1 - p_i) \right) + \sum_{i=1}^{p} \log(1 - p_i)$$

$$= \sum_{i=1}^{n} Y_i \left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} \right) - \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}}\right)$$

$$\frac{\partial \ell}{\partial \beta_m}(\beta_0, \dots, \beta_m) = \sum_{i=1}^{n} Y_i \cdot x_{im} - \sum_{i=1}^{n} \frac{x_{im} \cdot e^{\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}}}$$

$$= \sum_{i=1}^{n} Y_i x_{im} - x_{im} p_i$$

$$= \sum_{i=1}^{n} (Y_i - p_i) x_{im}$$

* **Poisson Regression** :

* **Example** : Neuroscience

Neurons in the central nervous system (CNS) transmit signals via a series of action potentials, or "spikes". The spiking of a single neuron may be measured by a system called microelectrode and it is a sequence of spikes over time is called a spike train. A simple and commonly used statistical model for a spike train is an Poisson point processes, which has the following properties.

For n time win draws of length Δ, letting Y_i denote the # of spikes generated by the neurons in the i^{th} time window, the random variables $Y_1, \dots, Y_n$ are indepentently distributed as $Y_i \sim \mathcal{P}(\lambda_i \Delta)$ where λ_i is controls the spiking rate in the i^{th} time window. For simplicity, we will assume $\Delta = 1$.

The spiking rate λ_i of a neureen may be influenced by external sensory stimuli present in this i^{th} window of time, for example, the intensity and pattern of light visible to the eye or texture of an object presented to the touch. To understant the effects of these sensory stimuli on the spiking rate of a particular neuren, we may perform an experiment that applies different stimuli in different windows of time and records the neural response. Encoding the stimuli applied in the i^{th} window of time by a set of p covariates $x_{i1}, \dots, x_{ip}$ a simple model for λ_i is given by :

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad \text{(Poisson log linear model)}$$

$$\lambda_i = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} = \mathbb{E}(Y_i \mid X_i) \quad , \quad \lambda_i > 0 \Rightarrow \log(\lambda_i) \in \mathbb{R}$$

Another assumption : Constant variance

* Interpretation of Coefficients :

$$\frac{\partial}{\partial x_{ij}} \mathbb{E}(Y_i \mid X_i) = \beta_j \cdot \exp(\beta_0 + \sum_{j=1}^{p} \beta_{ij} x_{ij}) = \beta_j \, \mathbb{E}(Y_i \mid X_j)$$

One-unit the j^{th} regressor leads to a change in the conditional mean by the amount $\beta_j \cdot \mathbb{E}(Y_i \mid X_j)$ (In LR $\frac{\partial}{\partial x_{ij}} \mathbb{E}(Y_i \mid X_i) = \beta_j$.

Practical Insights : Many of times we may have to transform

y_i (log y_i / y_i! a) before fitting poisson model.

* Estimation of the parameters :

$$L(\beta_0, \dots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} y_i^{\lambda_i}}{y_i!}$$

$$= \frac{e^{-\sum \lambda_i} \left(\prod y_i^{\lambda_i} \right)}{\prod_{i=1}^{n} y_i!}$$

$$\ell(\beta_0, \dots, \beta_p) = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \log(\lambda_i) - \sum_{i=1}^{n} \log(y_i!)$$

$$= -\sum_{i=1}^{n} e^{X\beta} + \sum_{i=1}^{n} y_i (X\beta) - \sum_{i=1}^{n} \log(y_i!)$$

$$\frac{\partial \ell(\beta_0, \dots, \beta_p)}{\partial \beta_j} = -\sum_{i=1}^{n} x_{ij} e^{X\beta} + \sum_{i=1}^{n} y_i x_{ij}$$

$$0 = \sum_{i=1}^{n} x_{ij} \left(y_i - e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right)$$ Newton Raphson

$$I_y(\beta) = -\mathbb{E}_\beta \left(\nabla^2 \ell(\beta) \right)$$

$$\frac{\partial^2 \ell(\beta_0, \dots, \beta_p)}{\partial \beta_l \partial \beta_j} = \sum_{i=1}^{n} x_{ij} x_{il} \left(-e^{\sum_{j=0}^{p} \beta_j x_{ij}} \right)$$

$$W = W(\beta) = \operatorname{diag}\left(e^{\sum_{j=0}^{p} \beta_j x_{1j}}, e^{\sum_{j=0}^{p} \beta_j x_{2j}}, \dots, e^{\sum_{j=0}^{p} \beta_j x_{nj}} \right)$$

$$\frac{\partial^2 \ell(\beta_0, \dots, \beta_p)}{\partial \beta_l \partial \beta_j} = -X^T W X = \nabla^2 \ell(\beta)$$

Hence, $I_Y(\beta) = X^T W X$ and thus $\hat{\beta} \sim N(\beta, (X^T W X)^{-1})$

Standard error of estimating $\hat{\beta}_j$: $\sqrt{(X^T \hat{W} X)^{-1}}$, $\hat{W}$ plugin estimate

* Modelling Assumption of Poisson Regression :

In real data, many of times the observed values of $Y_i \sim Pois(\lambda_i)$ is larger then the mean, this is called overdispersion.

If $Y_1, \ldots, Y_n$ are independent, $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ for each i (so the model for mean Y_i's is correct).

An alternative model for overdispered model is Negative Binomial ($X \sim NB(p, r)$ $E(X) = \frac{r(1-p)}{p}$ $Var(X) = \frac{r(1-p)}{p^2}$) so $Var(X) > E(X)$ as $0 \le p \le 1$ so it works for overdispersed.

When too many zero's in the data : Zero Inflated Data (Poisson / Negative Binomial) Theomphson Sampling (Tanujit Paper)

* Count time series arises in many domains :

1) Daily # of patients admitted to a hospital
2) The # of transactions of a given stock observed every minute

They are often overdispersed and highly autocorrelated, zero counts, possible models PR, NBR, ZIP, ZINB (ZIP zero count pois)

* Generalised Linear Model :

Let $Y_i \sim f(y \mid \theta_i)$ and $g(\theta_i) = \alpha + \beta x_i = x_i^T \beta$;

$g : \mathbb{R} \to \mathbb{R}$ is called the link function.

$$\begin{cases} \text{We get linear regression : } & g(Y_i) = \mathbb{E}(Y_i | X) = x_i^T \beta + \epsilon_i, \quad Y \sim N(\mu, \sigma^2) \\ \text{We get logistic regression : } & g(p) = \log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta, \quad Y \sim \text{Ber}(p) \\ \text{We get poisson regression : } & g(\lambda_i) = \log(\lambda_i) = x_i^T \beta, \quad Y \sim \mathcal{P}(\lambda) \end{cases}$$

Family members of GLM

* Exponential Family of Distribution :

$$f(y | \theta) = \exp(\theta y - A(\theta)) \cdot h(y)$$

$$f(y) = p^y (1-p)^{1-y} = (1-p) \left(\frac{p}{1-p}\right)^y$$

$$= \exp\left\{ y \underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{Natural Parameter = Link function}} + \log(1-p) \right\}$$

$$f(y) = \frac{e^{-\lambda} \lambda^Y}{Y!} = \exp\left\{ -\lambda + Y \underbrace{\log(\lambda)}_{\text{Natural Parameter}} - \log(Y_i) \right\}$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(Y-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{1}{2} \log(2\pi\sigma^2) + \frac{(Y-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left\{ \frac{y\mu}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \frac{\log(2\pi\sigma^2)}{2}\right) + \frac{y^2}{2\sigma^2} \right\}$$

* Cox Proportional Hazard Model :

* Exemple :

A clinical trial is performed to study the effect of a drug for maintaining prolonging remission induced by chemotherapy in the treatment of acute leukemia (Remission mean the disappearance of luhemic cells and other symptoms of the disease). For each i^{th} patient in the trial, let T_i denote the length of the length of the remission (or equivalently, the time unit recurrence of cancer), which we wish to model in terms of patient-specific covariates $x_{i1}, \dots, x_{ip}$.

Response : Time until the patient i comes back for treatment. Independent and time dependent but not time serves.

Features : Patient-specific variables.
Age - family history - previous treatment

* Modelling T_i as a continuous, positive-valued RV with CDF $F_i(t)$ and PDF $f_i(t) = F_i'(t)$. It is useful to think about the distribution of T_i in terms of the hazard function $\lambda_i(t)$ which represents the "instantaneous risk" of recurrence at time t:

$$\lambda_i(t) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}(T_i \leq t + \delta \mid T_i \geq t)$$

For small δ, the probability of the reccurance of cancer in the time window $[t, t+\delta]$ conditional on it not having occured up to time t is approximately $\delta \lambda_i(t)$. We also express this

$$\lambda_i(t) = \lim_{\delta \to 0} \frac{1}{\delta} \frac{\mathbb{P}(t \leq T_i \leq t+\delta)}{\mathbb{P}(T_i \geq t)} = \lim_{\delta \to 0} \frac{F_i(t+\delta) - F_i(t)}{1 - F_i(t)} = \frac{f_i(t)}{1 - F_i(t)}$$

Suppose that $T \sim \exp(\theta)$

$f_i(T) = \theta e^{-\theta t}$ $\qquad \lambda_i(t) = \dfrac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta$ constant!

$1 - F_i(t) = e^{-\theta t}$

We are assuming that the probability of reoccurrence is the same for every T and it is determined by θ. Hence, Cox PH model does not assume that $T \sim$ exp / parametric family instead models the hazard function of T_i as follows:

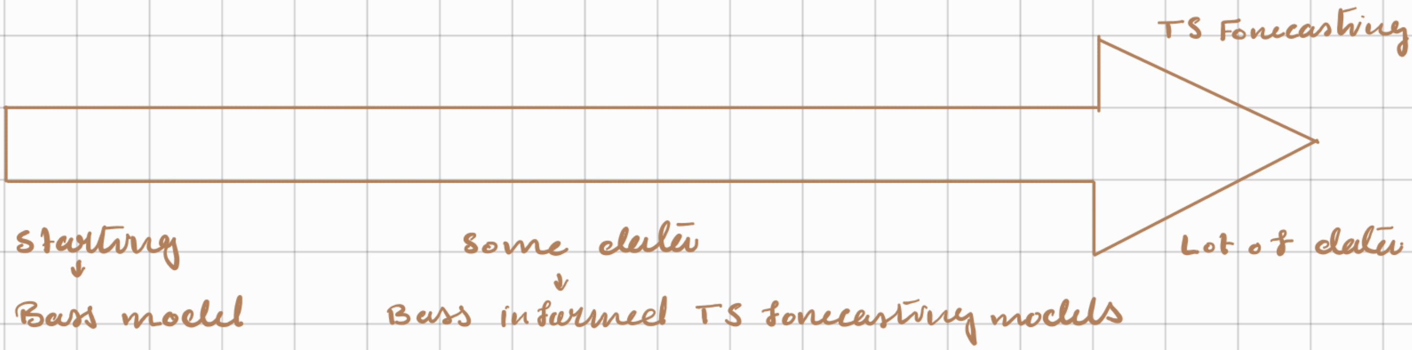
$$\lambda_i(t) = \lambda_0(t) \exp(\underset{\sim}{x}_i^t \underset{\sim}{\beta})$$

Proportional Hazard Model: $\dfrac{\lambda_i(t)}{\lambda_0(t)} \exp(\underset{\sim}{x}_i^t \underset{\sim}{\beta})$ where β_j's are unknown parameters determining the effects of the covariates on the remission length T_i and $\lambda_0(t)$ is the completely known baseline hazard function. $\lambda_0(t)$ controls the shape of the hazard function over time for all patients and $\exp(\underset{\sim}{x}_i^T \beta)$ controls the scale of the hazard function for each patient i. The ratio $\dfrac{\lambda_i(t)}{\lambda_j(t)}$ is constant over time (hence the name PH).

The baseline hazard when the coefficients of the predictors are zero.

* Bass Diffusion Model : Product Forecasting

Product forecasting is the science of predicting the degree of success of a new product will enjoy in a market place. To achieve this product forecasting model use several factors such as product awareness, distribution of a product, its price, fulfilling the market needs. Bass model is one way to forecast new product sales and technology. For example, forecasting IPHONE sales.



* Assumptions :

Bass diffusion model is used to forecast long-term trajectory for the sales pattern of new technologies and new durable products under two types of condition :

1) The firm has recently introduced the product / technology and has observed its sales for a few time period.

2) The firm hasn't introduced the product yet but it knows from existing products whose adoption pattern is known.

* The main idea of the model is that the adoption rate of a product comes from two main sources :

Bass Model

1) Innovation : Propensity of consumers to adopt the product independent of social influence to do so.

2) Imitation : The additional propensity to adopt the product

because others have adopted it (network effect).

Generalise Bass

3) Marketing Efforts : Ads, Social media, campaign, etc

* Example : Forecasting the evaluation of a start up company.

We define the cummulative probability of purchase of a product from time zero to time t by a single individual as $F(.)$. Then, the probability of purchase at time t $f(t) = \mathbb{P}(T=t) = F'(t)$.
The rate of purchase at time t given no purchase so far, is
$H(t) = \frac{f(t)}{1-F(t)}$ (hazard rate in survival analysis).

$\rightarrow$ Market size $= m$

Modelling $H(.)$ is just like modelling the adoption rate at time t. Therefore, Bass suggested that :

$$\frac{f(t)}{1-F(t)} = p + q\,F(t)$$

where p is the coefficient of innovation and q is the coefficient of imitation. p is the independent rate of a customer adapting the product and q is the imitation rate because it modulates the impact from the cummulative intensity of adoption $F(t)$.

If we can find p and q for a product, we can forecast its adoption over time and thereby generate a time path of sales.

$$\frac{f(t)}{1-F(t)} = \frac{F'(t)}{1-F(t)} = p + q\,F(t) \qquad (1)$$

$$\Rightarrow F'(t) = (1-F(t))\,(p+q\,F(t))$$

$$\Rightarrow \int \frac{1}{(1-F(t))(p+qF(t))}\, dF = \int dt = t + c_1$$

$$\frac{1}{(1-F(t))(p+qF(t))} = \frac{A}{1-F(t)} + \frac{B}{p+qF(t)}$$

$$\Rightarrow 1 = pA + qAF(t) + B - BF(t)$$
$$= (pA + B) + (qA - B)F(t)$$

$$\Rightarrow \begin{cases} pA + B = 1 \\ qA - B = 0 \end{cases} \Rightarrow \begin{cases} (p+q)A = 1 \\ B = qA \end{cases}$$

$$\Rightarrow \begin{cases} A = 1/(p+q) \\ B = q/(p+q) \end{cases}$$

$$\Rightarrow t + c_1 = \frac{1}{p+q} \int \frac{1}{1-F(t)}\, dF + \frac{q}{p+q} \int \frac{1}{p+qF(t)}\, dF$$
$$= \frac{-1}{p+q} \ln(1-F(t)) + \frac{1}{p+q} \ln(p+qF(t))$$

$$F(0) = 0 \Rightarrow c_1 = \frac{\ln(p)}{p+q}$$

Hence, $t + \dfrac{\ln(p)}{p+q} = \dfrac{\ln(p+qF(t)) - \ln(1-F(t))}{p+q} \Rightarrow$

$$t(p+q) + \ln(p) = \ln\left(\frac{p+qF(t)}{1-F(t)}\right) \Rightarrow$$

$$e^{t(p+q) + \ln(p)} = \frac{p+qF(t)}{1-F(t)} \Rightarrow$$

$$pe^{t(p+q)}(1-F(t)) = p + qF(t) \Rightarrow$$

$$pe^{t(p+q)} - p = F(t)(q + pe^{t(p+q)}) \Rightarrow$$

$$F(t) = \frac{p\left(e^{t(p+q)} - 1\right)}{q + p e^{t(p+q)}} \qquad (2)$$

$$f(t) = \frac{p(p+q)e^{t(p+q)}\left(q + p e^{t(p+q)}\right) - p^2(p+q)e^{t(p+q)}\left(e^{t(p+q)} - 1\right)}{\left(q + p e^{t(p+q)}\right)^2}$$

$$= \frac{pq(p+q)e^{t(p+q)} + p^2(p+q)e^{2t(p+q)} - p^2(p+q)e^{2t(p+q)} + p^2(p+q)e^{t(p+q)}}{\left(q + p e^{t(p+q)}\right)^2} \qquad (3)$$

$$= \frac{e^{t(p+q)}\, p\,(p+q)^2}{\left(q + p e^{t(p+q)}\right)^2}$$

If the target market is of size m, then the adoptions at time t is given by $m \cdot f(t) = s(t)$

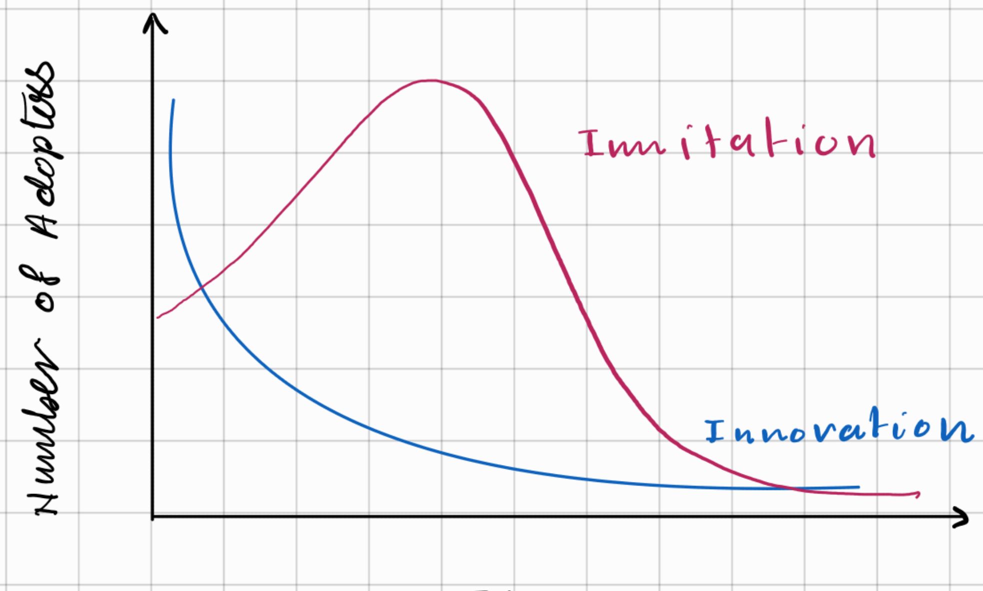
$q = 0 \Rightarrow$ exponential distribution

$p = 0 \Rightarrow$ logistic distribution

$$\frac{dF/dt}{1-F} = qF \Rightarrow \int \frac{1}{qF(1-F)}\, dF = \int dt = t + c_2$$

$$\Rightarrow \frac{1}{q} \int \frac{1}{F} - \frac{1}{1-F} = t + c_2$$

$$\Rightarrow \log(F) - \log(1-F) = qt + qc_2$$



Time

Generalised Bass Model : $\dfrac{f(t)}{1-F(t)} = (p + q F(t))\, x(t)$ (4)

$\downarrow$ marketing effort

* Model Calibration :

Q 1 : How to estimate the coefficients p and q ?

Given we have history of sales of the product, we can use it to fit the adoption curve.

Sales at any period one : $s(t) = m f(t)$

Cummulative sales at time t : $S(t) = m F(t)$

Substituting in 1 : $\dfrac{s(t)/m}{1 - S(t)/m} = p + q \dfrac{S(t)}{m} \Rightarrow$

$\beta_1 = pm$

$\beta_2 = q - p$

$\beta_3 = \dfrac{-q}{m}$

$$s(t) = pm + (q-p) S(t) - \frac{q}{m} (S(t))^2$$

$$= \beta_1 + \beta_2 S(t) + \beta_3 (S(t))^2$$

$$\beta_3 m^2 + \beta_2 m + \beta_1 = 0 \Rightarrow$$

$$m = \frac{-\beta_2 \pm \sqrt{\beta_2^2 - 4\beta_3\beta_1}}{2\beta_3}$$

$$p = \frac{\beta_1}{m} \qquad q = -m\beta_3$$

Sales peak : $t^* = \underset{t}{\arg\max}\,(f(t)) = \dfrac{\ln(q) - \ln(p)}{q+p}$

We will fit a linear model to compute $\beta_1, \beta_2, \beta_3$

$f(t) = (1 - F(t))(p + q F(t))$

$= p + (q-p) F(t) - q F(t)^2$

$$= \beta_1 + \beta_2 F(t) + \beta_3 F(t)^2$$